Send Orders for Reprints to reprints@benthamscience.net

 Current Psychiatry Reviews, **2014**, *10*, **330-347**

Collecting Information for Rating Global Assessment of Functioning (GAF): Sources of Information and Methods for Information Collection

I. H. Monrad Aas*

Research Unit, Division of Mental Health and Addiction, Vestfold Hospital Trust, PO Box 2267, 3103 Tönsberg, Norway

Abstract: *Introduction*: Global Assessment of Functioning (GAF) is an assessment instrument that is known worldwide. It is widely used for rating the severity of illness. Results from evaluations in psychiatry should characterize the patients. Rating of GAF is based on collected information. The aim of the study is to identify the factors involved in collecting information that is relevant for rating GAF, and gaps in knowledge where it is likely that further development would play a role for improved scoring.

Methods: A literature search was conducted with a combination of thorough hand search and search in the bibliographic databases PubMed, PsycINFO, Google Scholar, and Campbell Collaboration Library of Systematic Reviews.

Results: Collection of information for rating GAF depends on two fundamental factors: the sources of information and the methods for information collection. Sources of information are patients, informants, health personnel, medical records, letters of referral and police records about violence and substance abuse. Methods for information collection include the many different types of interview – unstructured, semi-structured, structured, interviews for Axis I and II disorders, semi-structured interviews for rating GAF, and interviews of informants – as well as instruments for rating symptoms and functioning, and observation. The different sources of information, and methods for collection, frequently result in inconsistencies in the information collected. The variation in collected information, and lack of a generally accepted algorithm for combining collected information, is likely to be important for rated GAF values, but there is a fundamental lack of knowledge about the degree of importance.

Conclusions: Research to improve GAF has not reached a high level. Rated GAF values are likely to be influenced by both the sources of information used and the methods employed for information collection, but the lack of research-based information about these influences is fundamental. Further development of GAF is feasible and proposals for this are presented.

Keywords: Collecting information, GAF, Global Assessment of Functioning, methodology, psychiatry, rating, severity of illness.

INTRODUCTION

In psychiatry, collecting information about patients is fundamental both for routine clinical practice and for research. The many instruments used for assessment in psychiatry differ in several ways, but they all depend on collecting information [1-3].

Global Assessment of Functioning (GAF) is an assessment instrument that rates severity of illness and is known worldwide [4,5]. An advantage of GAF lies in its simplicity [6]. Rating GAF is dependent on collecting information about psychological, social and occupational functioning. In its construction, GAF is an overall (global) measure of how patients are doing: it covers the range from positive mental health to severe psychopathology, and is intended to be a generic rather than a diagnosis-specific scoring system. Compared to diagnostic categories, GAF results in more comprehensive clinical descriptions [7]. Internationally, recorded GAF values can be a single value (only the most severe of symptom and functioning values is recorded) or recorded as separate values for symptoms (GAF-S) and functioning (GAF-F). For both GAF-S and GAF-F, there are 100 scoring possibilities (1–100) [4,5]. GAF has a wide range of applications. For example, it can be used in research and in the clinic to determine how well a treatment is working; improve the ability to link daily clinical work to empirical research; compare treatment results across diagnoses; allocate resources; and measure case-mix in mental health care.

There are problems with both reliability and validity associated with GAF. Reliability studies show the extreme 20% of raters can account for more than 50% of the spread of scores and there can be deviations of 20 points or more [8,9]. Overall reliability can be good, but it is lower in the routine clinical setting [9,10-16]. Concurrent validity

*Address correspondence to this author at the Research Unit, Division of Mental Health and Addiction, Vestfold Hospital Trust, PO Box 2267, 3103 Tönsberg, Norway; Tel: 47 33018000; Fax: 47 33018001;
E-mail: monrad.aas@siv.no

1875-6441/14 $58.00+.00

[10,12,17-30] and predictive validity [12,14,18,20,22,31-35] are more problematic. There are few empirical results for the sensitivity of GAF [5]. Further development of GAF will require work to improve its validity and reliability, and ensure good sensitivity and generic properties.

The nature and extent of conditions in psychiatry are rarely exposed to the same rigorous systematic scrutiny as their treatment [36]. More research on the design and evaluation of health measures is urgently needed [37]. Rigorous scientific study of assessment is required [38,39]. In general, assessors are likely to rate psychiatric impairments according to their own experience and attitudes, with resulting subjectivity in psychiatric evaluation [40]. Rating GAF is no exception [4,5]. There is evidence that different professions assign different GAF scores [41,42] and that the scores can be influenced by disagreement on the criteria for rating [9], lack of training [17], and problems related to the intrinsic properties of GAF itself [4]. It has been reported that the setting for investigation can explain some of the variability [30].

Instruments for assessment in psychiatry often do not describe or standardize a data-collection strategy [43]. However, both the sources used to make judgments and the manner of data collection from the sources are critical [43]. Evaluations should characterize patients, but most of the variability in diagnostics has been proposed to be a product of the evaluations and not the patients [44]. The development of effective interventions is slowed by problems with the evaluation instruments [43]. When raters of GAF assess on the basis of different information, different values may result [9]. An examination of GAF (with anchor points and examples of symptoms and functioning items) shows an obvious need for collecting information before rating. Information collection is a central element in the process for scoring GAF. Differences in information can explain some of the discrepancy in reliability studies for GAF [9]. However, little attention has been given to problems concerning collection of information. It has been proposed that all available information of importance for GAF should be examined [24]. The information available must, therefore, be sufficient for good overall judgement of symptoms and functioning.

The aim of this study is to identify factors concerning collection of information that is relevant for rating GAF, and gaps in knowledge concerning collection of this information. The gaps identified highlight areas where no, or little, research has been done and where it is likely that further development would play a role for improved scoring.

METHODS

Basic Literature Search

A literature review [45-47] was carried out. The search was conducted by both hand search and a search of bibliographic databases, in several steps:

a from previous work [4,5], the author had access to a considerable amount of literature about relevant issues, namely literature about GAF and other scoring systems, which also includes information about methodology;

b browsing through journals, which has been recommended as a useful first step before computer search [45] – each issue of a set of 17 journals for the period January 2000 to April 2011 (Acta Psychiatrica Scandinavica, American Journal of Psychiatry, Annals of General Psychiatry, Applied Psychological Measurement, Archives of General Psychiatry, BMC Psychiatry, British Journal of Psychiatry, Comprehensive Psychiatry, European Journal of Psychological Assessment, European Psychiatry, Evidence Based Mental Health, International Journal of Testing, Journal of Psychiatric Research, Psychiatric Services, Social Psychiatry and Psychiatric Epidemiology, The Journal of Clinical Psychiatry and The Psychiatrist);

c thorough hand search – after identification of publications by steps (a) and (b), their reference lists were hand searched for more literature and by, reading total publications, a search for citations to other studies was also conducted. Each time a relevant publication was identified, the same search for new literature was performed. After several rounds of such hand searching, new relevant references were not found and the search proceeded to steps (d) to (i);

d search for related articles to [5] on 19 August 2011 by using the 'related citations' option in PubMed – a total of 100 studies were identified;

e search for related articles to [4] on 22 August 2011 by using the 'related citations' option in PubMed – a total of 97 studies were identified;

f thorough hand search – after identification of publications by steps (d) and (e), their reference lists were hand searched for more literature and, by reading total publications, a search for citations to other studies was also conducted. Each time a relevant publication was identified, the same search for new literature was performed. After several rounds of such hand searching, new relevant references were not found and the search proceeded to steps (g) to (j);

g search in PubMed, which used experiences from research on search strategies [4]. A search was carried out for English language articles from the period January 1990 to July 2011. Search terms were: 'Global Assessment of Functioning OR GAF AND' combined with five search terms ('information', 'information AND source', 'informants', 'information AND collect*', 'observation') in five separate searches. A total of 308 studies were identified by this method. A further search was carried out for English language articles from the period January 1990 to November 2011. Search terms were: 'Global Assessment of Functioning OR GAF AND' combined with nine search terms ('guidelines', 'standard', 'reliability', 'validity', 'sensitivity', 'literature review', 'systematic review', 'psychometrics', 'methodology') in nine separate searches – a total of 2255 studies were identified by this method;

h a search in PsycINFO – this used experiences from research on search strategies [5]. A search was carried out for English language articles from the period January 1990 to 13 December 2011. Search terms were 'Global Assessment of Functioning OR GAF AND' combined

with seven search terms ('guidelines', 'instructions', 'information', 'interview', 'process AND rating', 'process AND scoring', 'methodology') in seven separate searches – a total of 1532 studies were identified by this method;

i possible missing publications remaining after steps (a) to (h) were controlled for by an Advanced Search in Google Scholar (for both books and articles) for the period from 1 January 2000 to the day the search was performed (21 December 2011). The search term 'Global Assessment of Functioning psychiatry' (used in one common search) identified 55,600 items (almost exclusively publications), and the first 1,000 were screened for relevance. Goggle Scholar gives information about the number of links to each publication (this is effectively a citation tracking with the most frequently cited publications listed first). The Google Scholar search did not identify any relevant studies that had not been already identified by steps (a) to (h);

j a search in The Campbell Collaboration Library of Systematic Reviews was carried out on 19 December 2011. The all-text searches were not limited to a specific time period. Five separate searches were performed (search terms: 'GAF', 'Global Assessment of Functioning', 'psychiatry systematic review', 'psychiatry literature review', 'psychiatry review'). However, this search identified no studies.

k the abstracts from steps (g) to (h) were screened, with the purpose of identifying literature relevant to collecting information for rating GAF. When this screening started, the researcher was experienced from reading literature from steps (a) to (f) and from two previous studies on GAF [4,5]. Abstracts were evaluated for inclusion by looking for information on the following issues in relation to GAF: sources of information, information collection, information-collection methods, strategy for information collection, rating process, methodology, psychometrics (studies with information on validity and reliability), interviews, interviews and methodology, interviews for rating GAF, biases for rating and diagnostics, patients and informants as information sources, guidelines, instructions, medical/mental health records, referrals, police records, observation, combining methods in psychiatry, computerized scoring, and modifications/changes made in the history of GAF. When the screening of abstracts was finished, selected publications were read in their entirety, but it became clear that most of the relevant literature had already been identified by steps (a) to (f);

l for the selected publications from step (k), the reference lists were hand searched for more literature. New publications that were relevant for inclusion were not found, and the literature search was complete. The Results section is based upon 125 references;

m the contribution of each selected publication to the knowledge base for the present study was summarized [45]. Emphasis was placed on points that were relevant for collecting information for rating GAF and analysis was performed to identify gaps in knowledge;

n the final set of selected publications is the reference list of the present study. Included publications are original research papers, books, chapters in books and articles.

RESULTS

The frontier of current knowledge and gaps in knowledge concerning collection of information for rating GAF were identified. The study shows that both the different sources of information and the different methods of information collection are likely to influence rated GAF values, but studies dealing specifically with these influences on GAF values were not found. The described frontier of current knowledge shows the progress that has been made by research.

Collecting Information for Rating GAF

Obtaining information for rating GAF is based on two fundamental factors: the source of the information and methods for information collection. Use of only one source and one method can result in incomplete and biased understanding [48-50]. The minimum standard for information collection has been proposed to be to read records of any previous treatment and to contact one relevant person [49]. The amount of information about the patient can play a role for the GAF score, but which information, from which sources, collected with which methods, for which patient categories, results in the best reliability is an open question. It is not necessarily true that more information results in better reliability. Lower reliability can result. This may be because of confusion as to how to incorporate different types of information into a GAF score [51]. The strategy for information collection strongly influences the type of information collected, but the choice of strategy is dependent on professional skills and experience [52].

Gap in knowledge: In general, the process of collecting information for GAF rating has been little studied and little is known about the importance of different processes of information collection for the GAF value generated. There is a scarcity of information about what the most important predictors of GAF are for different patient categories, i.e. which information, from which sources, obtained in which way, best predicts GAF. There is little research-based information about the minimum standard required for collection of information for different patient categories, i.e. information from which sources, collected with which methods, results in GAF values with acceptable reliability. Finally, little is known about which sources of information and which methods for information collection are required for different applications of GAF.

Sources of Information

Information for assessment can come from several sources: the patients, informants (for example family, friends, teachers, neighbours, work colleagues, employers), health personnel (for example general practitioners, nurses in home health care), medical records, letters of referral and police records about violence and substance abuse.

The 10 anchor points of GAF (with examples of symptoms and functioning items) have relevance for all the

above-mentioned sources of information. Research shows that the source of information exerts a prominent influence over judgements of the patient, but correlations between evaluations made with information from different sources can be low or moderate [48,53-55], and different sources of information can provide different pictures of the severity of illness [9]. In fact, every individual source of information can diverge substantially from every other potential source [48,50]. Also, different weight can be given to different sources of information; for example, some clinicians give greater weight to reports from clinical staff than to those from the patients themselves [44]. Obviously, use of different sources can result in different GAF scores, but use of several sources has the potential to corroborate, complement and correct results from individual sources [49]. When information from different sources represents different time periods, a change in symptoms over time can result in differences in recorded severity. For personality disorder, diagnoses using different sources of information can be substantially different [48,55,56].

Gap in knowledge: In the history of GAF, the importance of different sources of information for GAF scores has been little studied. It is not known which data source provides the best information. Little is known about how much weight should be given to the different information sources. The information-collection process can be different for different diagnostic groups, but there is little information about which sources of information are best for GAF rating for different diagnostic groups. It is not known whether different sources of information are necessary for reliable scores for different levels of severity. Low concordance between sources of information for the diagnosis does not necessarily mean the same as low concordance for GAF ratings, but little is known about this.

PATIENTS

Patients are an important source of information, but patients themselves can be unable to provide fully all types of information. In fact, agreement with a diagnosis based on a multimethod approach may be as low as 45–50% [48]. For example, people with schizophrenia and people with greater impairment in cognitive functioning cannot be expected to report everyday functioning accurately [57]. The self-report version of GAF is little used and has been little studied, but agreement with professionals' ratings can be good [58]. For personality disorders, agreement between self-report and interviews can be poor [59]. It should also be noted that self-reporting can underestimate behavioural problems [60].

Gap in knowledge: Patients themselves are important sources of information, but little analysis has been conducted on how important patients are for GAF values (compared to other sources of information). The importance may differ for Axis I and Axis II disorders, different diagnoses and the symptoms and functioning scales of GAF, but knowledge in this area is limited.

INFORMANTS

Information from different informants has only low to moderate concordance [48,55,56,60]. Scoring based on information from informants elicits the observer's perception of the patient [48]. When informants know the patient well, they have observed the patient over a longer time period in many settings. Knowledgeable informants can rate the patient as he or she usually is [61], but patient–informant correlations may be higher for friends than for family members [62]. Obtaining data from multiple informants has been considered optimal [56]. Some informants may be unwilling to report socially undesirable characteristics, and knowledge about a patient's previous mental illness may lead informants to pathologize behaviour more readily [55]. Discordance between informants is not necessarily connected to patient or informant demographic characteristics or how well informants know the patient [55], but research has shown that psychopathology in informants themselves can have an effect and it may be useful to take into account the psychiatric status of informants [62]. Nevertheless, informants' knowledge of the patient is important and should not be overlooked.

Gap in knowledge: Information from different informants may differ in its importance for GAF values, but knowledge in this area is limited. Data from different informants may complement and corroborate each other, but research on the importance of this for GAF values was not found.

PATIENTS VERSUS INFORMANTS

Agreement between patients and informants can be low [62], but patients' self-ratings and informants' ratings can agree well [61]. Informants can rate more psychopathology than patients are able to [61]. Informants (for example family members) have been considered more objective than patients, but this is not always the case [44,54]. For personality disorders, information from informants can provide other symptoms and substantially influence diagnoses [62,63], but for some personality disorders, interviews of patients can reveal more symptoms than informant interviews [54]. Concordance between a patient and informant can be better for some dimensional scores than for diagnoses [62].

Gap in knowledge: Information from both patients and informants can be important for GAF values, but the issue of having information from both has been little studied. It may be more important to have information from both for personality disorders than for other disorders. Data from patients and informants can correct and corroborate each other, but research on the importance of this for GAF values was not found.

HEALTH PERSONNEL

Health personnel can contribute information about patients. For example, general practitioners and nurses, who visit a patient at home, can contribute behavioural observations and information about the history of disease. When a clinician has observed symptoms and has access to symptoms reported by others [64], the question of reliability comes up. Personnel in psychiatry may view information from other health personnel as more credible, but the credibility is dependent on how well the other health personnel know the patients. A substantial number of

outpatients with schizophrenia, particularly those who are middle-aged and older, cannot identify a person who can report on their functioning [43]. Health personnel and other caregivers can then be informants [43]. Not all patients with schizophrenia are unable to identify informants willing to participate in structured interviews [57].

Gap in knowledge: Other health personnel can have a role to play in providing information about patients, i.e. by adding information about symptoms, functioning and the history of disease. How well they know the patient should be taken into consideration. Studies of the importance of information from other health personnel for GAF values were not found .

THE MEDICAL/ PSYCHIATRIC RECORD

The medical record can contain information about a patient's psychiatric history, symptoms, functioning, diagnosis, co-morbidity, assessments and tests performed with results, interviews with results, the presenting condition at different times, information from informants, letter(s) of referral, aetiology, treatment, prognosis and forensic history. It can be important to examine records of previous hospitalizations and outpatient treatments [49]. Clinicians vary remarkably in the depth and style of the information they record [44]. Psychopathology can be emphasized in the record and unusual or unique data are likely to be given disproportionate attention [44]. Clinicians rarely make notations about how questions are asked or whether the information recorded is a result of observation or is told by the patient or informant, but such factors are important [44]. A lack of adequate background information about the patient is a problem [3,43].

The GAF value chosen should be consistent with information in the medical record [40]. When positive psychology factors are well represented in descriptions of patients, more favourable GAF ratings can result [65]. The examiner should document which materials were reviewed for rating GAF [40]. When the medical record has information on the rationale for a given score, this eases the situation for clinicians taking over patients who have been treated by others [66]. How well the rater knows the patient should be specified [43]. Reporting of a GAF value in the psychiatric record can be followed by the time period (for example current or highest level in the past year or at admission and discharge) [67]. SCAN (Schedules for Assessment in Neuropsychiatry) includes a tool for scoring information from psychiatric records [68,69].

Gap in knowledge: Good medical records seem likely to contribute to more reliable rating of GAF, but empirical studies of the importance of records were not found . It is not known which factors in medical records are the most important for GAF values.

LETTERS OF REFERRAL

A referral can contain information about the patient's mental state at the time of writing, their history of mental health, treatment given, abnormal behaviour that may be concealed by the patient when interviewed by the psychiatrist, relevant points about their history of general health, work situation (for example unemployed and why), and relationships to family/friends/ work colleagues. From the referral, the clinician may develop, early on, a hypothesis about which diagnosis is right for a new patient [70-72]. The clinician can then choose instruments (for example the type of interview) according to the hypothesis.

Gap in knowledge: Information in referrals may influence GAF rating and the clinician's choice of instruments for information collection, but the present review could not identify any empirical work on the subject.

POLICE RECORDS

Police records can contain information that is important for rating GAF. The anchor points for the intervals 1–10 and 11–20 of the GAF scale include violence against others. Frequent shoplifting is one of the examples given for rating in the interval 41–50. Also, substance abuse can contribute to mental health, the choice of diagnosis, work and school functioning, relationships to others and crime.

Gap in knowledge: Information from police records may contribute to GAF values, but no empirical study on the issue was identified. It is not known whether the same information is reliably obtained from other sources or whether scoring based on other criteria is reliable.

Methods for Information Collection

The methods for information collection discussed here are interviews of patients and informants and observation of behaviour. A comprehensive and multidisciplinary approach, with both interviews and observation, has been found to be positively related to improving treatment outcomes [73].

THE INTERVIEW

The clinical interview is the psychiatrist's most important tool and has been called the core of all psychiatric research [49,64]. Many psychiatric interviews have been designed with the purpose of establishing a diagnosis [3,44,72,74-77]. However, there is a need for more multidimensional information about the patients than simply about diagnosis [53]. Psychological tests and different rating scales, including GAF, have a role to play.

The results from different diagnostic interviews, and even from different versions of the same interview, can be conflicting, for example they may vary significantly in estimates of prevalence [44,50,78,79]. This can be due to characteristics of the questions asked, change in symptoms, clinicians' bias, or deliberate attempts from the patient to distort information [44]. From methodology studies of the design of questionnaires, it is known that the wording and ordering of questions are important [80-82]. Even seemingly minor changes in the wording can have a major impact on scores [83]. Studies of instruments for psychiatry show the same importance of wording and ordering for the results [44,70,74,84-88]. Differences in information can also result when different people perform the same interview [89]. Training in administration of an interview has a role to play [59]. Standardized interviews represent an attempted to overcome these problems [89]. Obviously, different interviews

do not simply result in the same information. Information from general clinical interviews requires a 'translation' of answers into values on the GAF scale and this can be open to interpretation and some subjectivity [52,80,81,90].

OBSERVATION

Behavioural observation can be carried out in both healthcare settings and natural environments, and can be performance based in standardized, simulated environments [43,52,91]. Clinicians who rely exclusively on interviews are prone to incomplete understanding [48]. The clinician's early observation of the patient can shape their choice of assessment instruments (which dictates which information is collected). Nurses' behavioural observation of inpatients can contribute information. Natural environments can be, for example, the workplace, classroom or playground. Observation can be important for collected information.

Gap in knowledge: The many possibilities of interview design cannot be expected to result in the same information. Even when different people perform the same interview, differences in information may result. Studies of the importance of this for rated GAF values were not found . When both interviews and observation are used, they can supplement, correct and corroborate each other, but no information about the importance of this for GAF values was found.

UNSTRUCTURED, SEMI-STRUCTURED AND STRUCTURED INTERVIEWS

Structured patient interviews are much used in psychiatry today, but unstructured and semi-structured interviews are also used.

Unstructured Interviews

Unstructured interviews are not described in textbooks or overviews of different interviews, but are interviews created for individual clinical situations, and questions during the interview are changed according to the responses from the patient. Clinicians can perform unstructured interviews in different ways. Unstructured interviews can vary in focus, depth and duration, but be structured according to the clinician's understanding of relevant issues and knowledge of the patient [3,38,92]. The subjectivity of unstructured interviews is not entirely negative. They can be more revealing of what patients have in their minds and can cover eventualities not taken into consideration in structured interviews. However, subjectivity in the questions asked and the story told can mean the information gained is less balanced. Clinicians may focus on the presenting complaint and overlook other areas of functioning and symptoms [48]. For example, when clinicians fail to inquire about important aspects of psychopathology, the result can be significant problems for GAF rating. Information can be insufficient for a correct GAF value, but unstructured interviews can be combined with structured interviewing and questions can be added with the purpose of rating GAF.

Gap in knowledge: In the research on GAF, no information was found on whether unstructured interviews result in sufficient information for GAF scoring. It is not known whether a process consisting of performing structured and unstructured interviews first, then weighting the collected information, and finally assigning a GAF score, results in more reliable GAF values than a simpler process.

Semi-structured Interviews

A semi-structured interview is an interview with a guide about which questions should be asked, but questions can be open, i.e. without pre-decided answer alternatives. When semi-structured interviews are performed by clinicians, a better understanding of the constructs being assessed can result in relevant additional questions [38,79]. Examples of semi-structured interviews are; the SCID-I (Structured Clinical Interview for DSM [Diagnostic and Statistical Manual of Mental Disorders] Axis I), SCID -II (Structured Clinical Interview for DSM-IV Personality Disorders) and HDS (Hamilton Depression Scale) [44,93]. Information collected by semi-structured interviews can vary with the additional questions.

Semi-structured interviews can be more difficult for patients who are not used to formulating themselves or who have little education. Compared to fully structured interviews, the advantage can be greater variation and specificity in the answers. An overall impression of a patient can be obtained by performing a general diagnostic structured interview [94], but if too little information for a reliable GAF score is obtained, a semi-structured interview can be added [95]. In general, semi-structured interviews can be used to supplement fully structured interviews [38]. When the patient is well known to the clinician, a semi- structured interview for GAF scoring can be short. A longer semi-structured interview can be used when more questions are needed to obtain a reliable GAF rating and when greater accuracy about severity is important [95]. Compared to unstructured interviews, reliability has been found to be higher for semi-structured interviews [59]. Development of semi-structured interviews has enhanced the reliability of diagnosis of personality disorders [96].

Gap in knowledge: Research on the importance of semi-structured interviews for GAF rating is limited, but it should not be excluded that addition of a semi-structured interview to a structured interview can help in bridging the gap between information obtained by the structured interview and information that is relevant for reliable GAF scores.

Structured Interviews

In the fully structured interview, questions should be asked without any change in the way they are formulated, in the given ordering and without inserting new questions. The obtained information can be combined according to different algorithms, which are specific for each diagnosis [44,50,70, 77,97]. By reducing the variability in how interviews are conducted, the reliability of the information gathered can be improved [44,50,52,68,70,77]. Structured interviews are more valid than unstructured ones [44,86,92]. Several structured interviews are designed and used; examples are: the MINI (Mini-International Neuropsychiatric Interview), SCAN, CIDI (Composite International Diagnostic Interview) and DIS (Diagnostic Interview Schedule) [44,68-70,97-100].

Structured interviews can be different in their focus, the questions asked and the answer alternatives. Different structured interviews easily result in differences in the information collected [44].

Structured diagnostic instruments allow a systematic exploration of diagnostic criteria, with implicit information about severity [50] and consequently GAF values. The low flexibility of structured interviews can result in lack of coverage of some eventualities of what some patients have in their mind [3,44,48]. This may have consequences for GAF values. When the questions in an interview are less than complete, under-diagnosis can result [44] and this may well have consequences for the GAF value. An adherence to structured interviews that is too strict can limit understanding of the individual nature of mental disorders, by leaving the relationship between symptoms and functioning/ circumstances unexplored [44]. As GAF is about symptoms and functioning, lower understanding of the information needed for rating GAF can result. For reliable rating of GAF, a structured interview may require follow-up questions for clarification. Results from scoring with some instruments (instruments for depression, agitated behaviours, severity of general psychiatric symptoms, cognitive assessment in geriatric patients) show different correlation coefficients with GAF [101]. This could be explained by the differences in collected information. In anxiety disorder, suicidality can be significantly correlated to GAF ratings [102]. GAF scores can have higher correlations to global and total scores than to more detailed constructs [103]. As GAF is an overall assessment of mental health, this is not surprising. Research has not been adequate when it comes to comparing one structured diagnostic interview with another and examining whether different instruments are superior for certain diagnoses or clinical issues [79]. This also includes lack of information about the suitability of collected information for GAF rating. The history of GAF research has not made it possible to rank structured interviews according to their suitability for reliable GAF rating.

Gap in knowledge: It is not known whether different structured interviews, when used for the same patients, result in minimal or larger differences in GAF values. When the low flexibility of structured interviews results in lack of coverage of some eventualities, the collected information may be precise, but not necessarily the best for GAF rating. This issue has not been studied in any detail. The differences in information from different interviews may have different effects on the reliability of GAF-S and GAF-F, but research on the issue was not found. It is not known whether the different information from different interviews results in clear differences in the values for the single-scale GAF.

INTERVIEWS FOR AXIS I AND AXIS II DISORDERS

In DSM-IV, mental disorders are classified on Axis I ('clinical disorders') and Axis II ('personality disorders and intellectual disabilities') [44].

Axis I

Axis I diagnoses include, for example, major depressive episode, schizophrenic episode and panic attack [44]. Internationally, there are several instruments for diagnosis and screening of Axis I disorders, examples are: the SCID-I, MINI, SCAN, CIDI and DIS [76,99].

With the semi-structured interview, the SCID- I, information obtained in different ways can be used, such as information from observation, informants, and medical records [44,97,104]. In a sample of patients, those given a borderline diagnosis by the SCID and MCMI (Million Multiaxial Clinical Inventory) may not be the same patients [68]. The method of collection has a clear influence on the information collected. The MINI can be used by psychiatric organizations to meet the need for a short and accurate interview [76]. A look at the MINI interview shows a need for translating collected information into GAF values, but studies of the reliability of GAF values for different Axis I diagnoses were not found . The purpose of the SCAN is to assess, measure and classify the psychopathology and behaviour associated with a broad range of psychiatric disorders [68,104]. Information obtained by the SCAN has clear overlap with information of interest for rating GAF. For the CIDI, both reliability and validity are good [44,68,70,99,105]. When data from the CIDI are entered into the standard computer program, the output is a list of the criteria for diagnosis that are satisfied [70]. This information is relevant for rating GAF. In the DIS, the interviewee must answer 'yes' or 'no' to the questions asked and a computer program assigns the diagnosis [68,98]. When the patient answers affirmatively to questions about symptoms, he or she is asked about significant interferences with life or activities as a result of the symptoms [44,97,99]. Information about symptoms and functioning is of obvious relevance for GAF values.

In general, studies of structured interviews (like the CIDI and SCID-I) show that they can result in an incorrect diagnosis [48,70]. Data from different interviews are used to rate GAF [106].

Gap in knowledge: For Axis I diagnoses, it seems likely that the differences in information resulting from different interviews can result in different GAF values, but the issue has been little studied. The different Axis I interviews seem to vary in how well they are adapted to the symptom and functioning information needed for GAF rating. Information collected by the interviews needs to be translated into information used for GAF rating, but little is known about errors (and sources of errors) associated with the different interviews. Empirical research does not discriminate which Axis I interview results in the most reliable GAF values.

Axis II

Axis II disorders have an early onset and chronic course, and include, for example, antisocial personality disorder, avoidant personality disorder and borderline personality disorder [89]. The first interview for personality disorders was the SIDP (Structured Interview for DSM-III Personality Disorders). Today, several interviews for personality disorders exist, for example, the SCID- II, IPDE (International Personality Disorder Examination) and MMPI (Minnesota Multiphasic Personality Inventory) [39,44,68,93,96,97, 107,108].

In diagnoses of personality disorder, relationships, work, emotions and behaviour play a role for diagnosing, and the behaviour should have lasted for at least 5 years [44,59, 68,107,108]. For the same patients, interviews for different personality disorders can result in different diagnoses, but the reliability can also be quite good [44,59,68]. Functioning is important for GAF values. In general, patients with personality disorders are more impaired than those without [109]. The greater the total number of Axis II criteria met, the more severe the functional impairment [109]. Functional impairment attributed to some personality disorders may be more a function of their relationship to Axis I conditions than a direct impact on functioning itself [110].

Gap in knowledge: Research has provided little information about the suitability of information from different Axis II diagnostic interviews for GAF rating. Empirical research does not give information about which Axis II interview results in the most reliable GAF values. When different Axis II interviews result in differences in information, it is not known how different the GAF values become. GAF is intended to be a generic instrument, but it is not known whether the anchor points and examples make GAF equally well adapted to rating Axis II disorders as to rating Axis I disorders.

RATING OF SYMPTOMS

Symptoms are clinically important and the clinician should ask about the presence of symptoms. Batteries of single-item questions about symptoms are often referred to as symptom checklists [74]. Several symptom measures have been developed, for example the SCL-90-R (Symptom Checklist-90-Revised), GHQ (General Health Questionnaire) and MHI (Mental Health Inventory) [74,103,111].

Symptoms are important for GAF rating and deciding the GAF-S value. The current symptom anchor points of GAF were generally assigned in earlier stages of development of the scale. Much research on symptoms has been performed since then [4]. The correlation between GAF and number of symptoms can be significant [112]. GAF is significantly related to the SCL-90-R [65]. Correlation between GAF and the symptom score for schizophrenia can be high [103]. For patients with psychosis, GAF scores primarily reflect symptom severity [35]. In studies, data from the SCL-90-R, DIS and PSE (Present State Examination) and other interviews have been used to rate GAF [106], but symptom checklists are different and result in differences in collected information.

Both symptoms and clinical diagnoses have been proposed to be stronger predictors of GAF ratings than social or occupational functioning [14,102,113]. In the single-scale GAF, a weaker association with symptom scores than for the CGI (Clinical Global Impression – severity scale) may reflect the fact that only one value is recorded for GAF (the lower of symptom and functioning scores) [114]. Although the SCL-90-R is primarily a measure of symptom distress, it yields some supplement in assessing the extent of inter-personal dysfunction [115]. In GAF, reduced functioning values may well be related to change in psychiatric symptoms [116].

Gap in knowledge: High correlation between GAF and some symptom measures indicates that using symptom measures initially could improve GAF values (when rating symptom GAF is a problem), but there are few evaluations of this. Studies showing which symptom checklist results in the best information basis for rating GAF-S were not found . Based on newer symptom research, other symptoms could have been added to the GAF scale (as examples) and play a role for rating, for both Axis I and Axis II disorders, but there is no research showing that this would improve the generic properties, reliability and validity of GAF.

RATING OF FUNCTIONING

In psychiatry, assessment of functioning should be a part of comprehensive evaluation [117]. Functioning can mean social/familial relationships, vocational/educational functioning and self-care [118].

A large number of indices of functioning have been constructed [4]. Examples of functioning scales are: the WHODAS II (World Health Organization Disability Assessment Schedule II), SAS (Social Adjustment Scale), ICF (International Classification of Functioning), ADL (Activities of Daily Living), Katz Adjustment Scales, and PSP (Personal and Social Performance scale) [84,117,119-122]. There is no agreement on which functioning scale is best for different purposes [123]. Although the PSP has been proposed to have good psychometric properties [121], there is generally a lack of data on psychometric properties for functioning scales [117]. Both the WHODAS and SF-36 (Short-Form-36) are conceptually related to GAF, but the anchor points of GAF are not well specified [124]. Functioning scales are different, and result in differences in the information collected [123].

Functioning is one of the two fundamental characteristics of GAF. Treatment of schizophrenia requires resources, but there is no 'gold standard' for measuring functioning, i.e. there is a pressing need to develop good functioning scales for the disorder [123]. Patients with depression may well de-emphasize psychological symptoms and instead report physical symptoms, for example back problems with related problems of functioning [125]. For older patients, decline in cognitive functioning is associated with poorer performance of tasks of daily living [101]. In geropsychiatric patients, the overall level of functioning may be predicted by measures that assess overall cognitive and general psychiatric status [101]. The importance of such factors for GAF values is little researched. In the context of scales used to measure psychosocial functioning, there is generally poor assessment of psychometric properties [123].

It has been proposed that more attention should be paid to assessment concerning interaction between the patient and his or her context [1]. Functioning is related to contextual factors, but we are primarily interested in rating an individual's ability to function per se. For example, occupational functioning is not only related to mental health, but can be related to low education, physical health, stage in life and recent immigration. Large social networks are correlated with better functioning measured with GAF [126]. In GAF-F, the focus is on social and occupational

functioning, but it is not known whether social and occupational functioning contribute equally to the GAF-F score. It is possible that work plays a greater role [7].

It should be noted that different sources of information can give different information about functioning. Useful information can be obtained from informants [123]. Self-report may result in subjective scores [105], for example due to lack of insight and cognitive deficits [117]. Trained raters can rate functioning, but clinicians may be influenced by poor knowledge of the patients' day-to-day life [123]. For functioning, there is lack of standardization of the evaluation. For example, inter-centre differences in scoring of social functioning may be due to the lack of standardization [127].

Gap in knowledge: Compared to the problems with rating functioning in general, the issue of rating GAF-F is not well researched. While good information for rating GAF-F is lacking, the considerable international research on functioning has not yielded any information on which of the functioning scales could be used to obtain the best information for rating GAF, nor whether initially rating functioning by a functioning scale results in more reliable GAF-F values. It is not known whether adding information from informants improves the GAF-F values to a significant degree. Based on the research on functioning that is currently available, it would be helpful to add new examples of functioning to the GAF-F scale, but there is no research showing that this improves the generic properties, reliability and validity of GAF.

SYMPTOMS AND FUNCTIONING IN THE SINGLE-SCALE GAF

In the single-scale GAF, only the most severe of the symptom and functioning values are recorded. Use of only the most severe of functioning and symptoms will serve well in most cases, but recording only one value in the single-scale GAF has also been considered as incorrect [14,128]. The question of the comparability of single-scale GAF values is relevant. However, substantial differences between GAF-S and GAF-F values have been found in only 10% of cases [128]. There is no agreement between studies as to whether symptom or functioning scores in GAF are more severe [7,35,101,127-129]. The relative contribution of the GAF-S and GAF-F scores to the single-scale GAF score may differ across severity and diagnoses [35]. In general, symptom severity does not always correlate strongly with disability and only a small proportion of variability in disability is explained by any combination of clinical symptoms [130].

Gap in knowledge: The use of different sources of information and different methods for information collection result in differences in information, but it is not known whether this has a particular effect on single-scale GAF values.

SPECIALIZED SEMI-STRUCTURED INTERVIEW FOR RATING GAF

If collected information is difficult to interpret for GAF rating, a more specialized semi-structured GAF interview can be added. The anchor points and examples of the GAF scale give information about which questions such a semi-structured interview should include. The higher reliability of GAF at discharge can be explained by more information being available [19]. There are not many specialized interviews for scoring GAF, but in Norway a semi-structured interview for scoring has been developed [95].

Sometimes little information is needed to score GAF, particularly at the lower end of the scale. For example, knowledge of repeated suicide attempts requiring close supervision warrants a low score [94]. Scoring at the two highest 10-point intervals requires exclusion of the presence of symptoms, or the presence of minimal symptoms, and first-class or good functioning in a number of areas. Specialized semi-structured interviews for GAF rating can include questions about both inclusion and exclusion criteria for 10-point intervals [44].

GAF is intended to be a generic scoring system, but does not cover all psychiatry equally well. Generic properties mean that GAF should cover at least the most prevalent psychiatric disorders [131,132]. To cover all psychiatry, semi-structured interviews for different conditions can be developed and new examples included in the GAF scale (maybe also with changed anchor points). The GAF scale will then become better adapted to different conditions. Designing interviews for psychological assessment is not easy, and few interviews for scoring of GAF have been subjected to rigorous methodological analysis [86]. Interviewers can use their best judgement for scoring GAF, but there is still an inherent element of subjectivity. A scale could be designed to score within a decile [4]. In general, the design of scales can play a role for scorings [133].

Gap in knowledge: Research on specialized interviews for scoring of GAF is difficult to find. Little is known about the potential to improve GAF values by using specialized GAF interviews, or about the design of such interviews.

INFORMATION FROM INFORMANTS

Informants can be important sources of information [43]. There are documented methods for obtaining information from informants [48,54-56,60-63,134-137]. An example is the SCAN interview, which includes a tool for scoring information from informants [68,69]. The different interviews used for informants result in differences in the information collected. It is generally recommended that multiple diagnostic tools are used in research, to minimize error [79].

For evaluation of personality disorders, informants are particularly important. Both patients and informants can provide unique information [138], but there is sometimes poor agreement between the patient and informant [59]. Some researchers recommend that only informants should be used. Patients can lack insight and describe themselves in an idealized manner [59].

Asking informants for concrete data can help to avoid subjectivity – for example, 'how often did you observe what you describe (once a week, last time several months ago, etc)?'. Information obtained from multiple informants can increase the accuracy of the diagnostic estimate [139] and help with finding the correct GAF value. Correlations

between different informants' assessments can be variable [53,136,140]. Poor agreement between informants is a methodological problem [79]. However, disagreement among informants is not necessarily negative. It can be as useful as agreement, and help with GAF scoring.

If informants have their own mental health problems, objectivity can be reduced [43,141]. When patients do not have informants, health personnel and other caregivers can act as informants [43,57,123]. For employment history, interviews of work colleagues and employers can be added. Information from family members can be important and there are many studies of the reliability and validity of this type of information [43,49,123,142]. For severe disorders, information from family members is clearly more valid than for less severe conditions [56,142,143]. Family history reports are influenced by the characteristics of the informant [141] and can vary according to who the informants are. Parent–child agreement of symptom assessment can be low to moderate [50,144]. In general, information collected from informants can vary with the interview methods and characteristics of informants.

Gap in knowledge: Little is known about the importance for GAF values of different interview methods, which informants are used, informants' characteristics, which type of information is most reliable, characteristics of the relationship to the patient and which combination of patient and informant interviews is the best.

OBSERVATION

A reliance on interviews means a restriction in the range of signs and symptoms assessed [73]. As data gathered through direct observation of behaviour do not use an intermediary instrument for assessment, observation has been considered superior to other means of data acquisition [145].

During interviews, it is natural that the clinician observes the patient [49]. Signs and symptoms observed during the interview should be included in a comprehensive evaluation [49]. Recording of behaviour can begin with a narrative description of the client's difficulties and continue with considering the antecedents and consequences of the behavioural problem [52]. To measure behaviour, the included types of behaviour should be clearly defined [52], as well as, for example, their frequency and duration. For informal observation, subjectivity in what is recorded and in its interpretation is likely. Informal observations can be one key source of information [48].

Methods for scoring behaviour exist [43,146,147]. Scales have been developed for evaluation based on observation, for use in both inpatient departments and the community [73]. Examples of scales for observational assessment are: the NOSIE-30 (Nurse's Observation Scale for Inpatient Evaluation), AMPS (Assessment of Motor and Process Skills) and RAPP (Routine Assessment of Patient Progress) [73,148,149]. The RAPP correlates significantly with GAF [73]. Videotaping has been used in observation of behaviour [150]. The MSE (Mental Status Examination) is heterogeneous, as it applies to both standardized and non-standardized observations and enquiries of the patient [44]. The SCID interview can include information obtained by observation [44,97,104]. In the PSE, information observed by the interviewer during the interview can be used to score items [98]. Checklists are available for child behaviour – both parent-, teacher-, and self-report forms [151].

In ethnography, participant observation is a well-known method for observation [145,152]. It is also used in research within psychiatric institutions [153]. Staff in hospital wards can perform observation with the aim of preventing suicides, self-harm and violence [91,154]. Withdrawal is also an example of observable behaviour [155]. For inpatients, observation over longer periods of time can reveal symptoms. In controlled trials, studies of longitudinal observation can complement the results [147]. It is possible to standardize environments by simulating work situations or role-plays. When trained raters score behaviour, inaccuracy and bias are reduced [43].

Observation in the natural environment has advantages. The focus can be on real-life skills, but validity can be a problem [43]. Although informants can give information based on observation in the real-life situation, the evaluations can be subjective [43,156]. Patients' behaviour can be determined by situational factors [43,53]. Sometimes behaviours are difficult to observe because of their covert nature, for example, stealing and running away from home [53]. For children, agreement between parents and children can be better for behavioural problems than for symptoms [157]. It is likely that information collected by observation varies according to the study method and the person observing.

Gap in knowledge: Information collected by observation seems to be of importance for GAF values, but empirical research on the issue was not found . It is not known which method of observation is best and which observers are most important. The importance of observation for GAF scoring when patients have different characteristics (diagnosis, symptoms, co-morbidity, duration of illness, age, sex, etc) has not been explored. Research into the development of observation methods in simulated environments, which can predict behaviour in the natural environment, is limited.

COMBINING COLLECTED INFORMATION

No single method can be considered as the 'gold standard' for collection of information about patients [136]. Empirical considerations support the multimethod approach as a means to maximize the validity of assessment [48,55], but when several methods are used, rating GAF requires that the information collected is combined. Batteries of assessments can result in a wide range of conflicting information [48].

For diagnostics, it has been considered feasible to integrate conflicting data and assign a reliable diagnosis [56,141]. For GAF, the accuracy of the match between collected information and the anchor points (with examples) can vary. This accuracy may well be important for the reliability of diagnosis. Assigning GAF values can vary from easy to difficult. When all the information collected points in the direction of one and the same severity level, rating GAF becomes easy. However, it can be a challenge to decide

which GAF value is correct [51,60]. Discrepancies in information can be tackled in three ways: information from one source or one method may be considered as less credible and be given less weight; the clinician can try to seek clarifying information; or the clinician can assume the intermediate version to be more objective [44].

When the collected information is different in character, it must be synthesized to rate GAF [48,55]. When information is present about two or more symptoms, indicating the same GAF value, this could count more highly than one deviating symptom. Discrepancies between information from different sources can be taken up during a patient interview [107]. The rating process for GAF includes evaluating whether the collected information is sufficient for scoring. When required, supplementary assessments can be carried out to identify and describe signs and symptoms in terms of their type, extent and severity [49]. When the collected information makes several GAF values look relevant, the final value should not be more severe than the severest of the single values. For the weighting, it is important that the clinician has good understanding of the weaknesses and strengths of different psychiatric methods and sees the information collected in the light of this.

Information from the patient interview has been proposed to have greater weight than information from informants, but less so when the clinician lacks confidence in what the patient tells [56]. Information from informants who have had a lot of contact with the patient should count more highly than information from those with little contact [56]. Sometimes informant data can be especially convincing, because they use good examples or are supported by other evidence, for example treatment records [56]. Also, the question can be asked as to whether it is appropriate to reconcile informant and patient reports when they are significantly different [155]. All ratings can be considered to offer a unique perspective [60]. Incorporation of a patient's self-report and family history data may reduce measurement error [141].

If two clinicians disagree substantially about a GAF value, both cannot be right, but one or both can be wrong. Single clinician assessments can be replaced by team assessments. When teams evaluate the information available, a decision can be made by consensus. With substantial agreement, the estimate can be considered the final best estimate [56,139]. For assessments made by teams, it can be an advantage that team members have different experiences with the patient and evaluate the same experiences differently [107]. The team process can progress as follows: first a tentative value is presented and supporting evidence discussed [56], discrepancies in the information are discussed, and a final score is based on consensus [107]. When team consensus is used, based on several sources of information, the reliability of both Axis I and Axis II disorders can be good or excellent [56]. Consensus in a team is sensitive to team leadership and processes in the group [158]. When several ratings exist, it has also been proposed that the mean of all ratings could be used [15]. A different strategy for rating GAF can be to use the best estimate made by experienced clinicians reviewing all the available information [139]. Mean scores from expert panels have been proposed as reference norms [24]. The question can be asked as to whether a true value for GAF really exists.

A generally agreed approach to combining different types of information is lacking, but guidelines can be worked out [5,56]. Clearly, defined algorithms can be necessary to determine a GAF value. For diagnostics, an algorithm consisting of several steps has been used [159]. For rating GAF, the way in which to combine different information may well be different for different diagnostic groups. Further research is necessary concerning how to combine different types of information to obtain a GAF value [79]. A computer-based decision-tree process for GAF scoring also exists [160].

Gap in knowledge: All the different sources of information and methods for information collection may be of importance for GAF values, but it is difficult to rank sources and methods according to their importance. In the context of GAF rating, there is insufficient information to work out a good weighting system for conflicting data.

DISCUSSION

Methodology

This study can be defined as a systematic literature review [45-47]. Several important criteria for review articles are satisfied, such as defining the problem, informing the reader of the status of current research, identifying gaps and suggesting the next step [161].

Two previous studies [4,5] gave access to a considerable amount of relevant literature. The browsing through a number of journals and the first round of thorough hand search of literature, steps (b) and (c) in Methods, were carried out because it was considered that: (a) some relevant publications were likely not to be included in computerized databases; (b) the computerized databases may not contain relevant search terms in the titles and abstracts of studies; and (c) studies are not indexed with relevant terms. After the computerized searches, steps (d) and (e) in Methods, a new thorough hand search was added, i.e. step (f). All hand searches were stopped as soon as new relevant references were not found. A combination of searching reference lists and reading publications has been considered the most thorough way of hand searching [162]. PubMed includes more than 500 psychology-related journals [163], but as the search showed few publications dealing specifically with the study issue, the search was continued in PsycINFO. The PsycINFO search added little new knowledge. The citation tracking in Google Scholar is not completely reliable when it comes to listing the most frequently cited publications first, but screening of the first 1000 results represents a thorough Google Scholar search. For the searches in the computerized databases, relevant subject terms were developed. To avoid missing publications, two searches were performed in PubMed with different sets of search terms, a new set of search terms was used in the PsycINFO search, and the search in Google Scholar was different from the preceding searches. The search in The Campbell Collaboration Library of Systematic Reviews added no new studies. The searches in PubMed, PsycINFO, Google Scholar and The Campbell

Collaboration Library of Systematic Reviews are reproducible. The searches in PubMed, PsycINFO and Google Scholar, steps (g) to (i) in Methods, revealed that most of the publications were already identified by the previous searches (steps (a) to (f) in Methods). The step (l) shows that a stage had been reached where new perspectives could not be identified by reading more publications; this situation is described by the term 'saturation' from qualitative research. As no studies were found dealing specifically with the influence on GAF values of the different sources of information and the different methods of information collection, neither topic-specific inclusion criteria nor criteria for grouping included studies were developed. Present knowledge has not reached a level that makes this relevant. The very thorough literature search was motivated by the difficulty of finding studies with relevance for the study issue and the fact that the study issue implies a broad review of psychiatric methodology. Search in computerized databases only or only hand search would not have been enough. It is not considered likely that publications that could have changed the study findings were missed as a result of the search process. The design and conduct of the study protected against bias [47,164].

Collecting Information for Rating GAF

The present literature review has identified the state of knowledge about information collection for GAF rating. In the history of GAF research, there has been little focus on the requirements of the information-collection process. The lack of knowledge is fundamental. In work to develop a better GAF, a review of this kind can be valuable.

Much diagnostic variability has been proposed to be a product of evaluations and not the patients themselves [44], but collected data should characterize the patients. All sources of information and the different methods for information collection are of potential importance for GAF values. These values can vary with the sources, methods, different combinations of sources and methods, and different combinations of conflicting data. Little is known about how much influence different rating processes have on GAF values. As GAF is a widely used scoring system, its reliability and validity are important. If GAF is not good enough, different GAF values do not necessarily reflect a difference in the severity of illness.

The strategy for information collection readily influences the type of information collected. For a new patient, the clinician may have a hypothesis that is important for the choice of strategy for information collection (for example based on the referral). The choice of strategy can be dependent on professional skills and experience [52]. In recent decades, considerable progress has been made in standardizing how information can be gathered [165]. The structured interview is particularly relevant in this context. In daily practice, rating GAF may be based on information that is primarily collected for diagnostics. The aim with improved standardization of the rating process for GAF is to increase the comparability of scores (for example from different studies, across organizations, comparison of treatment results across diagnoses, before and after treatment values); help in assigning more accurate scores (choosing better between GAF points within the 10-point intervals); improve the possibility of linking daily clinical work to empirical research; make combination of scores in meta-analysis safer; and help in education and training of assessors (with implicit learning of how errors occur).

The present study raises the question of the accuracy required for GAF. If the mental health problem is not defined correctly for broader categories of patients, this will affect the ability to find the correct intervention. In the clinic, GAF is not used in isolation to characterize patients' mental health. The question is more how well GAF functions in the context of other information. The task is to find the right therapy for different variants and degrees of pathology [166].

Further Development for GAF

Studies of flaws and biases have led to the question as to how they could be avoided [165], but it is necessary to develop a clearly improved GAF [4,5]. The ultimate goal is to improve the quality of care by decreasing the proportion of inappropriate decisions. The total assessment process for GAF rating has been little studied [5]. When deciding which process to use, it is necessary to consider which process can best achieve the aims. The alternatives are: (1) work should be done to improve GAF; (2) to continue to use GAF as it is; (3) GAF should not be used. Perhaps the most common error is to dismiss existing scales too lightly [167]. It should not be considered impossible to improve GAF [4,5]. However, GAF is just one example of scoring systems. The alternative to a global system may be to use systems designed for specific populations, such as scoring systems for depression, etc. Diagnosis-specific scales can be useful with more specific information and may result in better reliability. For older patients, a modification of GAF may be necessary [101]. GAF is not suitable for rating learning disability [168].

Increased standardization of the scoring process is a proposal, but it is not a matter of course that more standardization results in better GAF scores and makes GAF equally reliable for all types of psychiatric populations. Increased standardization of information collection should not be destructive for the clinician–patient relationship. It is important for clinicians to earn the respect and trust of their patients [5]. A too-strict adherence to standardization of information collection can limit understanding of the patients and of the individual nature of their mental disorders [44]. The interviewer can become too 'protocol bound' [44]. Standardization of the information collection should not be too rigid, blocking new strategies for information collection and making the information collection time consuming. Simplicity of the total process of diagnosis and severity rating is important. It is not proven that more information results in better reliability for GAF. If a scoring process is made complex, errors are more likely to be introduced [44]. A new strategy for information collection should be implemented effectively at all sites, but implementation of new practices is not always successful [5]. In general, changing a scoring process is not necessarily an improvement.

Further development for GAF may take account of the following:

1 GAF values can vary with sources of information, methods for information collection, different combinations of sources and methods, and different ways of combining conflicting data, but there is a significant lack of knowledge about these factors. It seems more relevant to search for a simpler solution than to test out all possible combinations for all patient categories. Even the simplest information is lacking, i.e. researchers cannot tell clinicians which standardized diagnostic patient interviews give the best information for GAF rating for the conditions that are most common and most resource requiring for the health service. This is an obvious area for future research;

2 little research has been carried out on the Norwegian specialized semi-structured interview for rating GAF [95]. If future research can show this interview leads to more reliable GAF values, its use can be recommended;

3 properties of the GAF scale itself play a role in its reliability [4]. For example, inclusion of more examples (for the anchor points) may make GAF better adapted to the variation in collected information. Newer research, with symptom and functioning data, could give ideas about relevant examples. More examples do not necessarily improve reliability [4,169], but future research could clarify this;

4 it may be possible to develop better guidelines for rating and a manual with more information about GAF [5], but guidelines that require more work with rating raise the question of balance between simplicity and accuracy [76,170]. The DSM-IV instructions for rating GAF are incomplete and the Norwegian guidelines are more comprehensive [5]. Research has demonstrated that variation in guidelines influences the responses given by patients [2,5]. In the future, rules describing how to combine the information collected can be included in guidelines and form a recommendation for some patient interviews. Problems with the information collection can be included and discussed in a manual [5]. In the future, more research should be done on these issues. Current guidelines for scoring GAF are not the result of sophisticated development [4,5], but it can be difficult to avoid an element of subjectivity. Raters can be sceptical about giving a low score, because of negative labelling of clients [17]. The way in which scales are constructed can also give rise to social class bias. People from lower social class may suffer more from somatic illness and tend to express psychological disorders in somatic terms [87];

5 for GAF, computerization of scoring may well be the future [5] and result in values close to to a 'gold standard' determined by expert ratings. The use of electronic patient records makes new quality-assurance methods possible [5]. Computer programs for rating GAF exist, but have not been subjected to extensive research. There is no guarantee that computerization results in better scores [51,160], but newer knowledge can make computer programs more sophisticated [5]. Study of computerization of GAF scoring is an area for future research;

6 in general, changes in scoring processes are not necessarily improvements and new methods should only be widely adopted after studies of their reliability and validity. Work with a scoring instrument is not finished without a pilot study [80,171].

Additional Factors for Rating GAF

Two previous studies have focused on the properties of GAF and guidelines for rating it [4,5]. The present study has focused on collecting information for rating GAF, but other factors can also play a part for the choice of GAF value. Factors that have not been treated in the three studies include: characteristics of the interviewer (professional background, training and motivation); scoring by groups, individuals or independent experts; and cultural factors [4,5].

CONCLUSIONS

This literature review identifies the state of knowledge concerning problems with information collection for rating GAF. Despite the fact that rating GAF is based on collected information, there is a fundamental lack of knowledge about the requirements for information sources and methods for information collection. The study shows that:

a different sources and different methods for information collection lead to differences in the information collected, and these differences may influence GAF values, but the issue is little researched. In future studies, information about sources of information and methods for information collection should always be made clear;

b it is not known which sources, which methods, or which combination of sources and methods result in the best information base for rating GAF;

c the use of several sources and several methods does not necessarily give information pointing in the direction of one and the same GAF value, but the information can be conflicting;

d there is no established systematic approach to combining different types of information for GAF rating;

e the match between collected information and the anchor points (with examples) can vary, is not necessarily good, and can result in interpretation and subjectivity for rating GAF.

GAF values should reflect patients' severity of illness and not just the methodology used. Compared to the wide use of GAF and its potential in a number of applications, research on the process of scoring is far from advanced. Development of a well-functioning process of scoring has not been sufficiently guided by research. However, further development is possible and may result in a more reliable GAF.

FUNDING

The study was funded by the Vestfold Hospital Trust.

CONFLICT OF INTEREST

The author(s) confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

The author thanks Dr Penny Howes (Medical and Scientific Editing Service, UK) who provided assistance with the language.

ABBREVIATIONS

ADL	=	Activities of Daily Living
AMPS	=	Assessment of Motor and Process Skills
CGI	=	Clinical Global Impression – severity scale
CIDI	=	Composite International Diagnostic Interview
DIS	=	Diagnostic Interview Schedule
DSM	=	Diagnostic and Statistical Manual of Mental Disorders
GAF	=	Global Assessment of Functioning
GAF-F	=	GAF for functioning
GAF-S	=	GAF for symptoms
GHQ	=	General Health Questionnaire
HDS	=	Hamilton Depression Scale
ICF	=	International Classification of Functioning
IPDE	=	International Personality Disorder Examination
MCMI	=	Million Multiaxial Clinical Inventory
MHI	=	Mental Health Inventory
MINI	=	Mini-International Neuropsychiatric Interview
MMPI	=	Minnesota Multiphasic Personality Inventory
MSE	=	Mental Status Examination
NOISE	=	Nurse's Observation Scale for Inpatient Evaluation
PSE	=	Present State Examination
PSP	=	Personal and Social Performance scale
RAPP	=	Routine Assessment of Patient Progress
SAS	=	Social Adjustment Scale
SCAN	=	Schedules for Assessment in Neuropsychiatry
SCID-I	=	Structured Clinical Interview for DSM, Axis I
SCID-II	=	Structured Clinical Interview for DSM-IV Personality Disorders
SCL-90-R	=	Symptom Checklist-90-Revised
SF-36	=	Short-Form-36
SIDP	=	Structured Interview for DSM-III Personality Disorders
WHODAS II	=	World Health Organization Disability Assessment Schedule II

REFERENCES

[1] Fernández-Ballesteros R. Psychological assessment: future challenges and progresses. Eur Psychol 1999; 4: 248-262.
[2] Groth-Marnat G. Handbook of Psychological Assessment. New Jersey: John Wiley & Sons Inc. 2009.
[3] Mackinnon RA, Michels R, Buckley PJ. The psychiatric interview in clinical practice. 2nd ed. Washington DC: American Psychiatric Publishing Inc. 2006.
[4] Aas IHM. Global Assessment of Functioning (GAF): properties and frontier of current knowledge. Ann Gen Psychiatry 2010; 9: 20.
[5] Aas IHM. Guidelines for rating Global Assessment of Functioning (GAF). Ann Gen Psychiatry 2011; 10: 2.
[6] Yamauchi K, Ono Y, Ikegami N. The actual process of rating the Global Assessment of Functioning scale. Compr Psychiatry 2001; 42: 403-409.
[7] Skodol AE, Link BG, Shrout PE, Horwath E. The revision of Axis V in DSM-III-R: should symptoms have been included. Am J Psychiatry 1988; 145:825-9.
[8] Loevdahl H, Friis S. Routine evaluation of mental health: reliable information or worthless 'guesstimates'? Acta Psychiatr Scand 1996; 93:125-8.
[9] Vatnaland T, Vatnaland J, Friis S, Opjordsmoen S. Are GAF scores reliable in routine clinical use? Acta Psychiatr Scand 2007; 115: 326-30.
[10] Burlingame GM, Dunn TW, Chen S, *et al*. Selection of outcome assessment instruments for inpatients with severe and persistent mental illness. Psychiatr Serv 2005; 56: 444-51.
[11] Grootenboer EMV, Giltay EJ, Lern R van der, Veen Tineke van, Wee NJA van der, Zitman FG. Reliability and validity of the Global Assessment of Functioning scale in clinical outpatients with depressive disorders. J Eval Clin Pract 2012; 18: 502-507.
[12] Hilsenroth MJ, Ackerman SJ, Blagys MD, *et al*. Reliability and validity of DSM-IV axis V. Am J Psychiatry 2000; 157: 1858-63.
[13] Mezzich AC, Mezzich JE, Coffman GA. Reliability of DSM-III vs. DSM-II in cild psychopathology. J Am Academy of Child Psychiatry 1985; 24: 273-80.
[14] Moos R, McCoy L, Moos BS. Global Assessment of Functioning (GAF) ratings: determinants and role as predictors of one-year treatment outcomes. J Clin Psychol 2000; 56: 449-61.
[15] Söderberg P, Tungström S, Armelius BÅ. Reliability of Global Assessment of Functioning ratings made by clinical psychiatric staff. Psychiatr Serv 2005; 56: 434-8.
[16] Stewart AL, Greenfield S, Hays RD, *et al*. Functional status and well-being of patients with chronic conditions. Results from the Medical Outcomes Study. JAMA 1989; 262: 907-13.
[17] Bates LW, Lyons JA, Shaw JB. Effects of brief training on application of the global assessment of functioning scale. Psychol Rep 2002; 91: 999-1006.
[18] Goldman HH, Skodol AE, Lave TR. Revising axis V for DSM-IV: a review of measures of social functioning. Am J Psychiatry 1992; 149: 1148-56.
[19] Hall RCW. Global Assessment of Functioning. A modified scale. Psychosomatics 1995; 36: 267-75.
[20] Hay P, Katsikitis M, Begg J, Da Costa J, Blumenfeld N. A two-year follow-up study and prospective evaluation of the DSM-IV Axis V. Psychiatr Serv 2003; 54: 1028-30.
[21] Jones SH, Thorncroft G, Coffey M, Dung G. A brief mental health outcome scale reliability and validity of the Global Assessment of Functioning (GAF). Br J Psychiatry 1995; 166: 654-9.
[22] Niv N, Cohen AN, Sullivan G, Young A. The MIRECC Version of the Global assessment of Functioning scale: reliability and validity. Psychiatr Serv 2007; 58: 529-35.
[23] Patterson DA, Lee M-S. Field trial of the Global Assessment of Functioning Scale-Modified. Am J Psychiatry 1995; 152: 1386-8.
[24] Pedersen G, Hagtvedt KA, Karterud S. Generalizability studies of the Global Assessment of Functioning - split version. Compr Psychiatry 2007; 48: 88-94.
[25] Piersma HL, Boes JL. Agreement between patient self-report and clinician rating: concurrence between the BSI and the GAF among psychiatric inpatients. J Clin Psychol 1995; 51: 153-7.

[26] Rey JM, Stewart GW, Plapp JM, Bashir MR, Richards IN. Validity of Axis V of DSM-III and other measures of adaptive functioning. Acta Psychiatr Scand 1988; 77: 535-42.

[27] Robert P, Aubin V, Dumarcet M, Braccini T, Souetre E, Darcourt G. Effect of symptoms on the assessment of social functioning: comparison between Axis V of DSM III-R and the psychosocial aptitude rating scale. Eur Psychiatry 1991; 6: 67-71.

[28] Roy-Byrne P, Dagadakis C, Unutzer J, Ries R. Evidence for limited validity of the revised Global Assessment of Functioning Scale. Psychiatr Serv 1996; 47: 864-6.

[29] Salvi G, Leese M, Slade M. Routine use of mental health outcome assessments: choosing the measure. Br J Psychiatry 2005; 186: 144-52.

[30] Tungström S, Söderberg P, Armelius B-Å. Relationship between the Global Assessment of Functioning and other DSM Axes in routine clinical work. Psychiatr Serv 2005; 56: 439-43.

[31] Bacon SF, Collins MJ, Plake EV. Does the Global Assessment of Functioning assess functioning? J Ment Health Couns 2002; 24: 202-12.

[32] Fallmyr Ø, Repål A. Evaluering av GAF-skåring som del av Minste Basis Datasett [Evaluation of GAF-rating as part of Minimum Basis Dataset]. Tidsskrift for Norsk Psykologforening 2002; 39: 1118-19.

[33] Mellsop G, Peace K, Fernando T. Pre-admission adaptive functioning as a measure of prognosis in psychiatric inpatients. Aust N Z J Psychiatry 1987; 21: 539-44.

[34] Parker G, O'Donell M, Hadzi-Pavlovic D, Proberts M. Assessing outcome in community mental health patients: a comparative analysis of measures. Int J Soc Psychiatry 2002; 48: 11-19.

[35] Smith GN, Ehman TS, Flynn SW, *et al*. The assessment of symptom severity and functional impairment with DSM-IV Axis V. Psychiatr Serv 2011; 62:411-17.

[36] Godlee F. Who should define disease? BMJ 2011; 342: d2974.

[37] Martinez-Martin P. Composite rating scales. J Neurol Sci 2010; 289: 7-11.

[38] Klein DN, Dougherty LR, Olino TM. Toward guidelines for evidence-based assessment of depression in children and adolescents. J Clin Child Adolesc Psychol 2005; 34: 412-32.

[39] Lima EN, Stanley S, Kaboski B, *et al*. The incremental value of the MMPI-2: when does therapist access not enhance treatment outcome? Psychol Assess 2005; 17: 462-8.

[40] Ryu SG, Hong N, Jung HY, *et al*. Developing Korean Academy of Medical Sciences guideline for rating the impairment in mental and behavioural disorders: a comparative study of KNPA's new guidelines and AMA's 6th guides. J Korean Med Sci 2009; 24(Suppl 2): S338-S442.

[41] Laderman ER, Stein SM, Papanastassiou M. Flattened hierarchies and equality in clinical judgement. Ther Communities 1999; 20: 81-92.

[42] Schorre BEH, Vandvik IH. Global assessment of psychosocial functioning in child and adolescent psychiatry. A review of three unidimensional scales (CGAS, GAF, GAPD). Eur Child Adolesc Psychiatry 2004; 13: 273-86.

[43] Bellack AS, Green MF, Cook JA, *et al*. Assessment of community functioning in people with schizophrenia and other severe mental illness: a white paper based on an NIMH-sponsored workshop. Schizophr Bull 2007; 33: 805-22.

[44] Rogers R. Handbook of diagnostic and structured interviewing. New York: The Guilford Press 2001.

[45] Cooper H. Synthesizing research. A guide for literature reviews. Thousand Oaks: Sage Publications 1998.

[46] Hunt DL, McKibbon KA. Locating and appraising systematic reviews. Ann Intern Med 1997; 126: 532-8.

[47] Oxman AD. Systematic reviews: checklists for review articles. BMJ 1994; 309: 648-51.

[48] Meyer GJ, Finn SE, Eyde LD, *et al*. Psychological testing and psychological assessment. A review of evidence and issues. Am Psychol 2001; 56: 128-65.

[49] Mezzich E, Berganza CE, Cranach M von, *et al*. Essentials of the World Psychiatric Association's International Guidelines for Diagnostic Assessment (IGDA). Br J Psychiatry 2003; 182(Suppl 45): s37-s57.

[50] Renou S, Hergueta T, Flament M, *et al*. Entretiens diagnostiques structures en psychiatrie de l'enfant et de l'adolescent. Encephale 2004; 30: 122-34.

[51] Woldoff SB. Reliability of the Global Assessment of Functioning Scale. PhD thesis. Drexel University, Faculty of Drexel University 2004.

[52] Groth-Marnat G: Handbook of psychological assessment. New Jersey: John Wiley & Sons Inc. 2003.

[53] Achenbach TM, McConaughy SH, Howell C. Child/adolescent behavioral and emotional problems: implications of cross-informant correlations for situational specificity. Psychol Bull 1987; 101: 213-32.

[54] Dreessen L, Hildebrand M, Arntz A. Patient-informant concordance on the structured clinical interview for DSM-III-R personality disorders (SCID-II). J Pers Disord 1998; 12: 149-61.

[55] Ferro T, Klein D. Family history assessment of personality disorders: I. Concordance with direct interview and between pairs of informants. J Pers Disord 1997; 11: 123-36.

[56] Klein DN, Quimette PC, Kelly HS, Ferro T, Riso LP. Test-retest reliability of team consensus best-estimate diagnoses of Axis I and II disorders in a family study. Am J Psychiatry 1994; 151: 1043-7.

[57] Harvey PD. Update on cognition. Assessment of everyday functioning in schizophrenia. Innov Clin Neurosci 2011; 8: 21-4.

[58] Ramirez A, Ekselius L, Ramklint M. Axis V- Global Assessment of Functioning scale (GAF), further evaluation of the self-report version. Eur Psychiatry 2008; 23: 575-9.

[59] Zimmerman M. Diagnosing personality disorders. Arch Gen Psychiatry 1994; 51: 225-45.

[60] Lee SW, Elliot J, Barbour JD. A comparison of cross-informant behavior ratings in school-based diagnosis. Behav Disord 1994; 19: 87-97.

[61] Bagby RM, Rector NM, Bindseil K, Dickens SE, Levitan RD, Kennedy SH. Sef-report ratings and informants' ratings of personalities of depressed outpatients. Am J Psychiatry 1998; 155: 437-8.

[62] Riso LP, Klein DN, Anderson RL, Ouimette PC, Lizardy H. Concordance between patients and informants on the personality disorder examination. Am J Psychiatry 1994; 151: 568-73.

[63] Peselow ED, Sanfilipo MP, Fieve RR. Patients' and informants' reports of personality traits during and after major depression. J Abnorm Psychol 1994; 103: 819-24.

[64] Lingjærde O, Bech P, Malt U, Dencker SJ, Elgen K, Ahlfors UG. Skalaer for diagnostikk og sykdomsgradering ved psykiatriske tilstander. Del 1: Metodologiske aspekter [Scales for diagnosis and severity grading in psychiatry. Part 1: Aspects of methodology]. Nord J Psychiatry 1989; 43(Suppl 19): 1-39.

[65] Spalitto SV. Client strengths in assessment: examining the effects on clinical judgement. PhD thesis. University of Kansas, Department of Psychology and Research in Education and the Faculty of the Graduate School 2003.

[66] Rosse RB, Deutsch SI. Use of the Global Assessment of Functioning scale in the VHA: moving toward improved precision. Veterans Health Syst J 2000; 5: 50-8.

[67] American Psychiatric Association: Diagnostic and statistical manual of mental disorders, fourth edition, text revision (DSM-IV-TR). Washington DC: American Psychiatric Association 2000.

[68] Torgersen S. Det strukturerte psykiatriske intervjuet [The structured psychiatric interview]. In: Rönnestad MH, Lippe A von der, Eds. Det Kliniske Intervjuet [The Clinical Interview]. Oslo: Gyldendal Norsk Forlag AS 2002; pp. 525-46.

[69] World Health Organization. Schedules for clinical assessment in neuropsychiatry (SCAN). In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 44-8.

[70] Andrews G, Peters L. The psychometric properties of the Composite International Diagnostic Interview. Soc Psychiatry Psychiatr Epidemiol 1998; 33: 80-8.

[71] Godoy A, Gavino A. Information-gathering strategies in behavioural assessment. Eur J Psychol Assess 2003; 19: 204-209.

[72] Poole R, Higgo R. Psychiatric interviewing and assessment. Cambridge: Cambridge University Press 2006.

[73] Ehman TS, Higgs E, Smith GN, *et al*. Routine assessment of patient progress. A multiformat, change-sensitive nurses' instrument for assessing psychotic inpatients. Compr Psychiatry 1995; 36: 289-95.

[74] Bowling A. Measuring disease. A review of disease-specific quality of life measurement scales. Buckingham: Open University Press 1997.

[75] Herjanic B, Reich W. Development of a structured psychiatric interview for children: agreement between child and parent on individual symptoms. J Abnorm Child Psychol 1997; 25: 21-31.

[76] Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. The Mini-International Neuropsychiatric Interview (M.I.N.I.): The development and validation of a structured diagnostic interview for DSM-IV and ICD-10. J Clin Psychiatry 1998; 59(Suppl 20): 22-33.

[77] Brooks SJ, Kutcher S. Diagnosis and measurement of adolescent depression: a review of commonly utilized instruments. J Child Adolesc Psychopharmacol 2001; 11: 341-76.

[78] McClellan J, Bresnahan MA, Echeverria D, Knox SS, Susser E. Approaches to psychiatric assessment in epidemiological studies of children. J Epidemiol Community Health 2009; 63(Suppl. 1): i4-i14.

[79] McClellan JM, Werry JS. Introduction. Research psychiatric diagnostic interviews for children and adolescents. J Am Acad Child Adolesc Psychiatry 2000; 39: 19-27.

[80] Hansagi H, Allebeck P. Enkät och Intervju inom Hälso - och Sjukvård. Handbok för forskning och utvecklingsarbete [Questionnaire and Interview in Health and Healthcare]. Lund: Studentlitteratur 1994.

[81] Lydeard S. The questionnaire as a research tool. Fam Pract 1991; 8: 84-91.

[82] McColl E, Jacoby A, Thomas L, *et al*. Design and use of questionnaires: a review of best practise applicable to surveys of health service staff and patients. Health Technol Assess 2001; 5(31).

[83] Goodman R, Iervolino AC, Collishaw S, Pickles A, Maughan B. Seemingly minor changes to a questionnaire can make a big difference to mean scores: a cautionary tale. Soc Psychiatry Psychiatr Epidemiol 2007; 42: 322-7.

[84] Bowling A. Measuring health. A review of quality of life measurements scales. Buckingham: Open University Press 1993.

[85] Dunbar M, Ford G, Hunt K, Geoff D. Question wording effects in the assessment of global self-esteem. Eur J Psychol Assess 2000; 16: 13-19.

[86] Kici G , Westhoff K. Evaluation of requirements for the assessment and construction of interview guides in psychological assessment. Eur J Psychol Assess 2004; 20: 83-98.

[87] McDowell I, Newell C. Measuring health: a guide to rating scales and questionnaires. Oxford: Oxford University Press 1987.

[88] Thomson C. Introduction. In: Thomson C, Ed. The instruments of psychiatric research. Chichester: John Wiley & Sons 1989; pp. 1-17.

[89] Gelder M, Mayou R, Geddes J. Psychiatry. Oxford: Oxford University Press 2006.

[90] Del Greco L, Eastridge L, Marchand B, Szentveri K. Questionnaire development: 4. Preparation for analysis. Can Med Assoc J 1987; 136: 927-28.

[91] Manna M. Effectiveness of formal observation in inpatient psychiatry in preventing adverse outcomes: the state of the science. J Psychiatr Ment Health Nurs 2010; 17: 268-73.

[92] Casullo MM, Márquez MO. Interview (general). In: Fernandez-Ballesteros R, Ed. Encyclopedia of psychological assessment. Thousand Oaks: Sage 2003; pp. 481-7.

[93] Malt UF, Bech P, Dencker SJ, Elgen K, Ahlfors UG, Lingjærde O. Skalaer for diagnostikk og sykdomsgradering ved psykiatriske tilstander [Scales for diagnosis and severity grading in psychiatry]. Nordisk Psykiatrisk Tidsskrift 1990; 44: 98-238.

[94] Endicott J, Spitzer RL, Fleiss JL, Cohen J. The Global Assessment Scale; a procedure for measuring overall severity of psychiatric disturbance. Arch Gen Psychiatry 1976; 33: 766-71.

[95] Karterud S. Semistrukturert Intervju for GAF-Vurdering [Semi-structured Interview for GAF rating] Available from: www.dagbehandlingsnettverk.no/kvalsikr/skjemaer/GAF_intervju.pdf

[96] Farmer RF, Chapman AL. Evaluation of DSM-IV personality disorder criteria as assessed by the structured clinical interview for DSM-IV personality disorders. Compr Psychiatry 2002; 43: 285-300.

[97] Bech P, Malt UF, Dencker SJ, *et al*. Scales for assessment of diagnosis and severity of mental disorders. Acta Psychiatr Scand 1993; 87(Suppl 372).

[98] Hasin DS, Skodol AE. Standardized diagnostic interviews for psychiatric research. In: Thomson C, Ed. The instruments of psychiatric research. Chichester: John Wiley & Sons 1989; pp. 19-57.

[99] Kobak KA, Skodol AE, Bender DS. Diagnostic measures for adults. In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 35-60.

[100] Sheehan DV, Lecrubier Y, Sheehan KH, *et al*. MINI International Neuropsychiatric Interview (MINI). In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 48-51.

[101] Whitney JA, Kunik ME, Milonari V, Lopez FG, Karner T. Psychological predictors and discharge Global Assessment of Functioning Scale scores for geropsychiatric inpatients. Aging Ment Health 2004; 8: 505-13.

[102] Schwartz RC, Pete-Brown TD. Construct validity of the Global Assessment of Functioning scale for clients with anxiety disorder. Psychol Rep 2003; 92: 548-50.

[103] Haro M, Kamath SA, Ochoa S, *et al*. The Clinical Global Impression-Schizophrenia scale: a simple instrument to measure the diversity of symptoms present in schizophrenia. Acta Psychiatr Scand 2003; 107(Suppl 416): 16-23.

[104] First MB, Spitzer RL, Gibbon M, Williams JBW. Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I). In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 40-3.

[105] Pull CB, Wittchen HU. CIDI, SCAN and IPDE: structured diagnostic interviews for ICD 10 and DSM III-R. Eur Psychiatry 1991; 6: 277-85.

[106] Beiser M, Fleming JAE, Iacono WG, Lin T. Refining the diagnosis of schizophreniform disorder. Am J Psychiatry 1988; 145: 695-700.

[107] Pedersen G. Diagnostiske intervjuer og GAF-vurdering. [Diagnostic Interviews and GAF-rating]. In: Karterud S, Urnes Ø, Pedersen G, Eds. Personlighetsforstyrrelser. Forståelse, evaluering, kombinert gruppebehandling. [Personality disorders. Understanding, evaluation, combined group treatment]. Oslo: Pax Forlag 2000; pp. 237-9.

[108] Paap MCS, Meijer RR, Bebber J van, *et al*. A study of the dimensionality and measurement precision of the SCL-90-R using item response theory. Int J Methods Psychiatr Res 2011; 20: e39-e55.

[109] Nakao K, Gunderson JG, Phillips KA, *et al*. Functional impairement in personality disorders. J Pers Disord 1992; 6: 24-33.

[110] Hong JP, Samuels J, Bienvenu J, *et al*. The longitudinal relationship between personality disorder dimensions and global functioning in a community-residing population. Psychol Med 2005; 35: 891-5.

[111] Delman HM, Robinson DG, Kimmelblatt CA, McCormack J. General psychiatric symptoms measures. In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 61-82.

[112] Lange C, Heuft G. Die Beeinträchtigungsschwere in der psychosomatischen und psychiatrischen Qualitätssicherung: Global Assessment of Functioning Scale (GAF) vs. Beeinträchtingungs-Scwere-Score (BSS). Z Psychosom Med Psychother 2002; 48: 256-69.

[113] Moos RH, Nichol AC, Moos BS. Global Assessment of Functioning ratings and the allocation and outcomes of mental health services. Psychiatr Serv 2002; 53: 730-7.

[114] Goldman M, DeQuardo JR, Tandon R, Taylor SF, Jibson M. Symptom correlates of global measures of severity in schizophrenia. Compr Psychiatry 1999; 40: 458-61.

[115] Pedersen G, Karterud S. Using measures from the SCL-90-R to screen for personality disorders. Personal Ment Health 2010; 4: 121-32.

[116] Jovanovic AA, Gasic MJ, Ivkovic, Milanovic S, Damjanovic A. Reliability and validity of DSM-IV Axis V scales in a clinical sample of veterans with posttraumatic stress disorder. Psychiatria Danubina 2008; 20: 286-300.

[117] Figueira ML, Brissos S. Measuring psychosocial outcomes in schizophrenia patients. Curr Opin Psychiatry 2011; 24: 91-9.

[118] Goodman SH, Sewell DR, Cooley EL, Leavitt N. Assessing levels of adaptive functioning: the role functioning scale. Community Ment Health J 1993; 29: 119-31.

[119] Brissos S, Palhavã F, Marques JG, *et al*. The Portuguese version of the Personal and Social Performance scale (PSP): reliability, validity, and relationship with cognitive measures in hospitalized

and community schizophrenia patients. Soc Psychiatry Psychiatr Epidemiol 2012; 47: 1077-86.
[120] Dickerson FB. Assessing clinical outcomes: The community functioning of persons with serious mental illness. Psychiatr Serv 1997; 48: 897-902.
[121] Morosini P-L, Magliano L, Brambilla L, Ugolini S, Pioli R. Development, reliability and acceptability of a new version of the DSM-IV Social and Occupational Assessment Scale (SOFAS) to assess routine social functioning. Acta Psychiatr Scand 2000; 101: 323-9.
[122] Williams JBW. Mental health status, functioning, and disability measures. In: Rush AJ, First MB, Blacker D, Eds. Handbook of psychiatric measures. Washington DC: American Psychiatric Publishing Inc. 2008; pp. 83-105.
[123] Brissos S, Molodynski A, Dias VV, Figueira ML. The importance of measuring psychosocial functioning in schizophrenia. Ann Gen Psychiatry 2011; 10: 18.
[124] Dimsdale JE, Jeste DV, Patterson TL. Beyond the Global Assessment of Functioning: learning from Virginia Apgar. Psychosomatics 2010; 51: 515-19.
[125] Greden JF. Physical symptoms of depression: Unmet needs. J Clin Psychiatry 2003; 64(Suppl 7): 5-11.
[126] Westermeyer J, Neider J. Social networks and psychopathology among substance abusers. Am J Psychiatry 1988; 145: 1265-9.
[127] Gaite L, Vázquez-Barquero JL, Herrán A, *et al*. Main determinants of Global Assessment of Functioning score in schizophrenia: a European multicenter study. Compr Psychiatry 2005; 46: 440-6.
[128] Pedersen G, Karterud S. The symptom and function dimensions of the Global Assessment of Functioning (GAF) scale. Compr Psychiatry 2012; 53: 292-8.
[129] Brekke JS: An examination of the relationship of three outcome scales in schizophrenia. J Nerv Ment Disord 1992; 180: 162-7.
[130] Fountoulakis KN, Lekka E, Koudi E, Chouvarda I, Deligiannis A, Maglaveras N. Development of the Global Disability Scale (GloDis): prelimiary results. Ann Gen Psychiatry 2012; 11: 14.
[131] Bijl RV, Ravelli A, Zessen G van. Prevalence of psychiatric disorder in the general population: results of the Netherlands mental health survey and incidence study. Soc Psychiatry Psychiatr Epidemiol 1998; 33: 587-95.
[132] Kessler RC, McGonagle KA, Zhao S, *et al*. Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States. Arch Gen Psychiatry 1994; 51: 8-19.
[133] Rammstedt B, Krebs D. Does response scale format affect the answering of personality scales? Assessing the big five dimensions of personality with different response scales in a dependent sample. Eur J Psychol Assess 2007; 23: 32-8.
[134] Andreasen NC, Rice J, Endicott J, Reich T, Coryell W. The family history approach to diagnosis. Arch Gen Psychiatry 1986; 43: 421-9.
[135] Gerahon ES, Guroff JJ. Information from relatives. Diagnosis of affective disorders. Arch Gen Psychiatry 1984; 41: 173-80.
[136] Salbach-Andrae H, Lenz K, Lehmkuhl U. Patterns of agreement among parent, teacher and youth ratings in a referred sample. Eur Psychiatry 2009; 24: 345-51.
[137] Weissman MM, Wickramararatne P, Adams P, Wolk S, Verdeli H, Olfson M. Brief screening for family psychiatric history. Arch Gen Psychiatry 2000; 57: 675-82.
[138] Klein DN. Patients' versus informants' reports of personality disorders in predicting 7½-year outcome in outpatients with depressive disorders. Psychol Assess 2003; 15: 216-22.
[139] Leckman JF, Sholomskas D, Thomson D, Balanger A, Weissman MM. Best estimate of lifetime psychiatric diagnosis. A methodological study. Arch Gen Psychiatry 1982; 39: 879-83.
[140] Baruch G, Fearon P, Gerber A. Emotional and behavioural problems in adolescents/young adults receiving treatment at a community-based psychotherapy centre for young people: a preliminary study of the correspondence among adolescent/young adult and significant other reports. Br J Med Psychol 1999; 72: 251-65.
[141] Kendler KS, Prescott CA, Jacobsen K, Myers J, Neale MC. The joint analysis of personal interview and family history diagnoses: evidence for validity of diagnosis and increased heritability estimates Psychol Med 2002; 32: 829-42.
[142] Hardt J, Franke P. Validity, reliability, and objectivity of the family history method in psychiatry: a meta analysis. Eur Psychiatry 2007; 22: 49-58.
[143] Fogelson DL, Neuchterlein KH, Asarnov RF, Payne DL, Subotnik KL. Validity of the family history method for diagnosing schizophrenia, schizophrenia-related psychoses, and schizophrenia-spectrum personality disorders in first-degree relatives of schizophrenia probands. Schizophr Res 2004; 68: 309-17.
[144] Edelbrock C, Costello AJ, Dulcan MK, Conover CN, Kala R. Parent-child agreement on child psychiatric symptoms assessed *via* structured interview. J Child Psychol Psychiatr 1986; 27: 181-90.
[145] Suen HK, Ary D. Analyzing quantitative behavioral observation data. New Jersey: Lawrence Erlbaum Associates Inc. 1989.
[146] Alevizos P, DeRisi W, Liberman R, Eckman T, Callahan E. The behavior observation instrument: a method of direct observation for program evaluation. J Appl Behav Anal 1978; 11: 243-57.
[147] Lecrubier Y, Perry R, Milligan G, Leeuwenkamp O, Morlock R. Physician observations of positive and negative symptoms of schizophrenia: a multinational cross-sectional survey. Eur Psychiatry 2007; 22: 371-9.
[148] Bouwens SFM, Heugten CM van, Aalten P, *et al*. Relationship between measures of dementia severity and observation of daily life functioning as measured with the Assessment of Motor and Process Skills. Dement Geriatr Cogn Disord 2008; 25: 81-7.
[149] Palmstierna T, Wistedt B. Staff observation aggression scale, SOAS: Presentation and evaluation. Acta Psychiatr Scand 1987; 76: 657-63.
[150] Chung K-M, Reavis S, Mosconi M, Drewry J, Matthews T, Tassé MJ. Peer-mediated social skills training program for young children with high-functioning autism. Res Dev Disabil 2007; 28: 423-36.
[151] Conners CK, Barkley RA. Rating scales and checklists for child psychopharmacology. Psychopharmacol Bull 1985; 21: 809-43.
[152] Reeves S, Kuper A, Hodges BD. Qualitative research methodologies: ethnography. BMJ 2008; 337: a1020.
[153] Oeye C, Bjelland AK, Skorpen A. Doing participant observation in a psychiatric hospital - research ethics resumed. Soc Sci Med 2007; 65: 2296-306.
[154] Muralidharan S, Fenton M. Containment strategies for people with serious mental illness. Cochrane Database Syst Rev 2006; (3): CD002084.
[155] Angold A, Weissman MM, John K, *et al*. Parent and child reports of depressive symptoms in children at low and high risk of depression. J Child Psychol Psychiatry 1987; 28: 901-15.
[156] Sripada BN, Jobe TH, Helgason CM. From fuzzy logic toward plurimonism: the science of active and empathetic observation. IEEE Trans Syst Man Cybernt B Cybern 2005; 35: 1328-39.
[157] Hodges K, Wotring J. The role of monitoring outcomes in initiating implementation of evidence-based treatments at the state level. Psychiatr Serv 2004; 55: 396-400.
[158] Aas IHM: The organizational challenge for health care from telemedicine and e-health. [monograph on the internet]. Oslo: The Work Research Institute 2007 [cited 2013 January 11]. Available from: http://www.afi.no/stream_file.asp?iEntityId=2088
[159] Pincus HA, Wise T, First MB, McQueen LE. DSM-IV Primary care Version: an opportunity for general hospital and consultation-liason psychiatrists? Gen Hosp Psychiatry 1995; 17: 324-5.
[160] Harel TZ, Smith DW, Rowles JM. A comparison of psychiatrists' clinical-impression-based and social workers' computer-generated GAF scores. Psychiatr Serv 2002; 53: 340-2.
[161] Bern DJ. Writing a review article for Psychological Bulletin. Psychol Bull 1995; 118: 172-7.
[162] Conn VC, Isaramalai S, Rath S, Jantarakupt P, Wadhawan R, Dash Y. Beyond MEDLINE for literature searches. J Nurs Scholarsh 2003; 35: 177-82.
[163] Arnold SJ, Bender VF, Brown SA. A review and comparison of psychology-related electronic resources. J Electron Res Med Libs 2006; 3: 61-79.
[164] Egger M, Jüni P, Bartlett C, Holenstein F, Sterne J. How important are comprehensive literature searches and the assessment of trial quality in systematic reviews? Health Technol Assess 2003; 7(1).
[165] Bruyn EEJ de. Assessment process. In: Fernandez-Ballesteros R, Ed. Encyclopedia of psychological assessment. Thousand Oaks: Sage 2003; pp. 93-7.
[166] Widiger TA, Clark LE. Toward DSM-V and the classification of psychopathology. Psychol Bull 2000; 126: 946-63.
[167] Streiner DL, Norman GR. Health measurement scales. A practical guide to their development and use. Oxford: Oxford University Press 1994.

[168] Oliver P, Cooray S, Tyrer P, Cicchetti D. Use of the Global Assessment of Functioning scale in learning disability. Br J Psychiatry 2003; 182(Suppl 44): s32-s35.

[169] Rey JM, Starling J, Weaver C, Dossetor DR, Plapp JM. Inter-rater reliability of global assessment of functioning in a clinical setting. J Child Psychol Psychiatry 1995; 36: 787-92.

[170] Kerstin M, Hornke LF. Improving the quality of proficiency assessment: the german standardization approach. Psychol Sci 2006; 48: 85-98.

[171] Del Greco L, Walop W. Questionnaire development: 1. Formulation. Can Med Assoc J 1987; 136: 583-5.

Received: January 11, 2013 Revised: June 18, 2013 Accepted: December 06, 2013